DOCUMENT RESUME

ED 392 822 TM 024 471

AUTHOR Harwell, Michael
TITLE An Empirical Study of the Hedges (1982) Homogeneity Test.
PUB DATE Apr 95
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22,1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; *Meta Analysis; Monte Carlo Methods; *Sample Size; Scores; *Statistical Distributions
IDENTIFIERS Fixed Effects; *Homogeneity Tests; *Power (Statistics); Type I Errors

ABSTRACT

The test of homogeneity developed by L. V. Hedges (1982) for the fixed effects model is frequently used in quantitative meta-analyses to test whether effect sizes are equal. Despite its widespread use, evidence of the behavior of this test for the less-than-ideal case of small study sample sizes paired with large numbers of studies is contradictory, and its behavior for nonnormal score distributions in primary studies is an open question. The results of a Monte Carlo study indicated that the Type I error rate and power of the homogeneity test were insensitive to skewed score distributions, but were very sensitive to smaller study sample sizes paired with larger numbers of studies. These findings extend earlier results and help to clarify the statistical behavior of the homogeneity test. Specifically, the pairing of small study sample sizes with large numbers of studies tends to produce conservative Type I error rates for the homogeneity test and underestimates its power, increasing the likelihood of Type II errors. (Contains 2 tables and 23 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 392 822

An Empirical Study of the Hedges (1982) Homogeneity Test

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MICHAEL HARWELL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Michael Harwell

University of Pittsburgh

April 1995

Paper presented at the annual meeting of the American Educational Research Association, San Francisco. Correspondence concerning this paper should be directed to Michael Harwell, 5H33 Forbes Quad, University of Pittsburgh, PGH, PA 15260

TM 024471

BEST COPY AVAILABLE

Abstract

Hedges' (1982) test of homogeneity for the fixed effects model is frequently used in quantitative meta-analyses to test whether effect sizes are equal. Despite its widespread use, evidence of the behavior of this test for the less-than-ideal case of small study sample sizes paired with large numbers of studies is contradictory, and its behavior for nonnormal score distributions in primary studies is an open question. The results of a Monte Carlo study indicated that the Type I error rate and power of the homogeneity test were insensitive to skewed score distributions, but were very sensitive to smaller study sample sizes paired with larger numbers of studies. These findings extend earlier results and help to clarify the statistical behavior of the homogeneity test. Specifically, the pairing of small study sample sizes with large numbers of studies tends to produce conservative Type I error rates for the homogeneity test and underestimates its power, increasing the likelihood of Type II errors.

An Empirical Study of the Hedges (1982) Homogeneity Test

The homogeneity test for fixed effects models proposed by Hedges (1982) provides a vehicle to model variability among effect sizes that has been widely used in meta-analysis.[1] For example, the journal **Psychological Bulletin** published 43 quantitative meta-analyses during the seven-year period from 1988-1994, 23 of which (53%) employed Hedges' homogeneity (Q) test. The genesis of this paper was the informal observation that published meta-analyses reporting Q tests (including the 23 meta-analyses using the Q test in **Psychological Bulletin** articles) rarely comment on whether the assumptions underlying this test are tenable, specifically, that the scores in primary studies are independently and normally distributed with a common variance and that the large sample properties of the test hold for small study sample sizes. Wolf (1990), among others, has expressed similar concerns.

The limited attention paid to assessing the assumptions of the Q test in published meta-analyses may be attributable to editorial policy devoted to minimizing the length of a paper, or to meta-analysts counting on the insensitivity of the Q test to assumption violations. For example, meta-analysts may know that the assumptions of the two sample t-test must technically be satisfied to ensure the validity of the Q test but be unconcerned with violations of the normality and equal variance assumptions because of the abundant analytic (e.g., Gayen, 1949; Srivastava, 1959) and simulation evidence (Harwell, Rubinstein, Hayes, & Olds, 1992; Sawilosky & Blair, 1992) documenting the robustness of the t-test. (The sensitivity of the t-test to dependencies is well documented). Of course, it is also possible that the lack of attention paid to assumption violations is the result of simple neglect.

---

[1]Many tests of homogeneity are available (c.f., Alexander, Scozzaro, & Borodkin, 1989); however, only the test of homogeneity of effect sizes representing standardized mean differences is considered.

Regardless of why published meta-analyses have paid little attention to assumptions of the Q test, it is not at all clear that the well-documented robustness of the t-test to nonnormality and small sample sizes is transmitted to the Q test. This study examined the effect of nonnormal score distributions in primary studies and their interaction with small study sample size and large numbers of studies on the Q test.

Why Does the Power of the Q Test Matter?

The Q test provides evidence of the adequacy of the model specified through the null hypothesis (Shadish & Haddock, 1994, p. 267). Chang (1992) described potential problems with using the Q to test for the adequacy of an explanatory model. In contrast to hypothesis testing in many primary studies, meta-analysts are often content to retain the tested null hypothesis since this suggests that whatever model is being tested adequately characterizes the variation in the effect sizes. In many ways, the use of the Q test in meta-analyses mimics the use of stepwise multiple regression procedures in which a nonsignificant result is often used as evidence that the regression model at the previous step is adequate for explaining variation in the outcomes. Chang suggested that meta-analysts should be especially concerned about the likelihood of Type II errors (i.e., retention of a false null hypothesis) since a Q test which was under-powered would lead to an unacceptably high probability of wrongly concluding that the model fits the data.

Chang described another reason why the power of the Q test is a concern. Retention of the homogeneity hypothesis is often followed by pooling the sample effect sizes and testing whether the weighted average effect size differs from 0. This two-stage procedure breaks down if the Q test of homogeneity has an unacceptable high probability of a Type II error since Ho would be retained too often, meaning that the results of the test of the average effect size in the

second stage may be misleading. Thus, factors which increase the probability of a Type II error beyond acceptable levels for the Q test are a special concern.

The Q Test

Consider a collection of effect sizes for i = 1, 2, ..., k studies involving two independent groups. The effect size is defined as

$$\delta_i = (\mu_i^E - \mu_i^C)/\sigma_i \qquad (1)$$

where δ_i is the population effect size for the ith study, μ_i^E and μ_i^C are population means on some metric variable Y, and σ is the standard deviation assumed to be common to both populations. (The notation used in Hedges & Olkin (1985) is followed). The unbiased estimator of δ_i is approximately

$$d_i = \left[\frac{\bar{Y}_i^E - \bar{Y}_i^C}{s}\right]\left[1 - \frac{3}{4N_i - 9}\right] \qquad (2)$$

where $\bar{Y}^E$ is the sample mean of the experimental group, $\bar{Y}^C$ is the sample mean of the control group, s is the sample pooled within-groups standard deviation (assuming $\sigma_E = \sigma_C$), and N_i is the total sample size for the ith study. The d statistic in equation (2) is also the minimum variance estimator of δ_i and is distributed as a noncentral t meaning that hypothesis testing involving the d_i takes on the usual assumptions of the two sample t-test for independent means.

Hedges (1982) used the fact that the large- sample distribution of d_i is normal to construct tests of the homogeneity of the δ_i. If the group sample sizes within a study n_i^E and n_i^C increase at the same rate, then, asymptotically, $d_i \sim N(\delta_i, \sigma_{\delta_i}^2)$, where σ_δ^2 is approximated by

$$\sigma_d^2 = \frac{N_i}{n_i^E n_i^C} + \frac{d_i^2}{2N_i} \qquad (3)$$

(Hedges & Olkin, 1985, p. 86).

The hypothesis Ho: $\delta_1 = \delta_2 = ... = \delta_k$ is tested with the statistic

$$Q = \Sigma_i (d_i - d_+)^2/\sigma_{d_i}^2 \qquad (4)$$

where d_+ is an average of the d_i weighted by $[\sigma_d^2]^{-1}$. Under Ho, Q is asymptotically distributed as a central chi-square variable with k-1 degrees of freedom. As noted above, retention of Ho is typically followed by pooling the d_i and testing the weighted average d_+ against 0, i.e., Ho: $\delta_+ = 0$. Hedges and Olkin (1985, p. 112) showed that $d_+ \sim N(\delta_+, \sigma^2_{\delta+})$.

If the δ_i are not equal then Q has a noncentral chi-square distribution k-1 degrees of freedom and noncentrality parameter (Chang, 1992):

$$\lambda = \Sigma_i \frac{(\delta_i - \delta_+)^2}{\sigma^2_{\delta_i}} \quad . \qquad (5)$$

Review of the Literature

Box (1953) noted that the insensitivity of the Type I error rate and power of a test to assumption violations is an important consideration in evaluating the test. The widespread use of the Q test suggests that its Type I error rate and power have been widely studied under realistic conditions (e.g., small sample sizes and skewed score distributions in primary studies). Surprisingly, this does not appear to be the case.

Wolf (1990) pointed out that published meta-analyses have paid little attention to the consequences of failing to satisfy the underlying assumptions of various meta-analytic tests and that little work has been done to evaluate the effect of assumption violations [A notable exception has been the development of robust and nonparametric effect size estimators.] Rosenthal and Rubin (1982) noted that the behavior of tests of homogeneity was not well

understood, a comment echoed by Chang (1992), who indicated that little was known of the power of the Q test for realistic settings such as small numbers of studies and small sample sizes within primary studies.

Hedges and Olkin (1985, p. 125) reported the results of a Monte Carlo study of the fit between the chi-square distribution and the distribution of Q when the δ_i were equal (but not zero). Their results indicated that, conditional on the score distributions being normally distributed with a common population variance, k = 5 resulted in slightly conservative Type I error rates for $N_i = 20$ and somewhat less conservative values for $N_i = 100$. In all cases, sample sizes within studies were equal. However, whether the between-study sample sizes were equal or unequal appeared to have no effect on Type I error rates. Hedges and Olkin (1985, p. 124) indicated that, on the whole, the Q test appears to be slightly conservative, which suggests that the probability of a Type II error may be slightly higher than might be desired, and that the large-sample approximation to the Q distribution improves as δ_i and N_i increase.

Chang (1992) performed a Monte Carlo study to examine the Type I error and power of the Q test which appears to be the most exhaustive investigation available. Chang began by surveying approximately 60 published meta-analyses for the period 1985-1990 for guidance in selecting simulation factors and their values. Chang investigated varying numbers of studies (k = 2, 5, 10, 30), sample size pairings (e.g., $n_i^E = n_i^C = 10$, $n_i^E = 10$, $n_i^C = 20$), and various noncentrality patterns, including (a) All but one of the k effect sizes were the same $\delta_1 = \ldots = \delta_{k-1} = 0$ and $\delta_k = .1, .25, .5, .75, 1$, (b) All but two effect sizes were the same $\delta_1 = \ldots = \delta_{k-2} = 0$ and δ_{k-1}, $\delta_k = .1, .25, .5, .75, 1$, (c) Simulating three clusters of δ values, 0, ...,0; δ,...,δ, and 2δ,...,2δ. Chang simulated t statistics for the primary studies under the assumption that the raw scores were independently and normally distributed with a common variance, and summarized her findings by comparing theoretical and power curves with goodness-of-fit tests and by using analysis of

variance and regression procedures to model variation in the empirical proportions of rejections as a function of study characteristics.

Chang drew the following conclusions after comparing the empirical and theoretical power values (a) The maximum discrepancies occured when larger numbers of studies (e.g., k = 30) were paired with smaller study sample sizes (e.g., N_i = 20). Because Chang did not report the actual power values it is difficult to judge the magnitude of the discrepancies, although there is evidence that most of the discrepancies were less than .2. (b) The fit between empirical and theoretical power curves was quite good for larger N_i (e.g., 60) regardless of the value of k (c) Whether primary studies had equal or unequal sample sizes did not have much effect on how closely empirical power curves matched their theoretical counterparts (d) Type of noncentrality pattern appeared to affect the fit between empirical and theoretical power values, particularily as k increased, although larger study sample sizes tended to mitigate this effect. The pattern of one extreme effect size and the rest equal produced the most discrepancies with the empirical power values typically exceeding the theoretical values. These results support the observation of Fleiss and Gross (1991) that a single study (i.e., a single effect size) may exert a powerful effect on the meta-analytic results. Chang also reported that the small N_i, large k pairing produced inflated Type I error rates, a finding which conflicts somewhat with that reported in Hedges and Olkin (1985, p. 125), although the latter study was limited to k = 2, 5. Inflated Type I error rates may explain why Chang's empirical power values exceeded theoretical power values for these same conditions. For other N_i and k pairings, Chang's Type I error results were generally consistent with those reported in Hedges and Olkin (1985, p. 125).

In short, the available evidence suggests that, conditional on the scores in primary studies being independently and normally distributed with a common variance, the Type I error rate and power of the Q test are close to theoretical values except for the case in which small N_i are

paired with larger k. It is important to note that Chang's survey provided evidence that this particular pairing does occur in published meta-analyses and, thus, should be of some concern for meta-analyses using the Q test under these conditions.

Methodology

Ideally the effect of assumption violations on the Type I error rate and power of the Q test would be studied using analytic methods; unfortunately, such solutions are quite difficult or impossible. As a substitute, a Monte Carlo study was performed to address the following research question: What are the effects of nonnormal score distributions on the Type I error rate and power of the Q test for varying study sample sizes and numbers of studies (assuming one effect size per study)? In all cases the data were homoscedastic.

Simulation Factors

The factors and their values selected for the Monte Carlo study reflect those of Chang (1992) and Hedges and Olkin (1985, p. 125). The design of the Monte Carlo study was a four-factor, fully-crossed factorial involving k, N_i, δ_i and type of score distribution.

Recall that Chang found little effect on Type I error rates and power for normally distributed scores for large N_i. Since skewness appears to play an important role in the behavior of tests of location parameters (c.f., Harwell, et al., 1992), three increasingly skewed distributions were simulated and identified by their skewness (γ_1) and kurtosis (γ_2). These were moderately skewed and leptokurtic chi-square distributions with $\nu = 8$ degrees of freedom ($\gamma_1 = 1$, $\gamma_2 = 3$), skewed and leptokurtic ($\gamma_1 = 1.5$, $\gamma_2 = 5$), and a chi-square with $\nu = 2$ ($\gamma_1 = 2$, $\gamma_2 = 6$). Data for a normal distribution ($\gamma_1 = \gamma_2 = 0$) were also simulated.

Other factors in the Monte Carlo study focused on Chang's findings for small sample sizes paired with large numbers of studies. The numbers of studies modeled were k = 5, 10, and 30, with group sample sizes of 5,5, 10,10, and 20,20. Unequal sample sizes were not included because of Chang's finding that equal and unequal sample sizes for the within or between study cases appeared to have the same effect on the Q test.

Only the noncentrality pattern studied by Chang which produced the most dramatic effect on power values, i.e., $\delta_1 = \ldots = \delta_{k-1} = 0$ and $\delta_k = 0, .5, 1, 1.5$, was studied. The noncentrality effect was created by adding the appropriate δ value to each score in the targeted group. For example, for k = 5, $\delta_1 = \ldots = \delta_4 = 0$, $\delta_5 = .5$ was added to each raw score in group 1 in study 5.

Following the recommendations of Naylor, Balintfy, Burdick, and Chu (1968), Hoaglin and Andrews (1975), Lewis and Orav (1989), and others that Monte Carlo studies should be treated as statistical sampling experiments subject to the same guidelines as empirical studies, the empirical Type I error rates and power for the Q test were analyzed using inferential procedures. This enabled the contribution of sampling error to be evaluated and the magnitude of significant effects to be estimated.

Data Generation

The data generation was done using a Gateway 4DX-33 486 microcomputer. All programming was done in FORTRAN IV supplemented by locally-written subroutines. The following process was performed to generate data: (a) N_i standard normal deviates were simulated using a random number generator given in **Numerical Recipes** (Press, Flannery, Teukolsky, & Vetterling, 1986), which were transformed to the the specified nonnormal form following the method of Fleishman (1978). These values were then assigned to one of two groups. (b) Constants equal to the specified δ values were added to scores in the target groups

to create the desired noncentrality pattern. (c) The d_i were calculated using equation (2) and σ_d^2 using equation (3), (d) Steps (a)-(c) were repeated k times, simulating the results of a single meta-analysis with k effect sizes. (e) The Q statistic was computed for the k effect sizes using equation (4) and compared to the appropriate central chi-square critical value at the α = .01, .05, and .10 levels of significance. (f) Steps (a)-(e) were repeated 2000 times (The same number of replications employed by Hedges (1982) and Chang (1992)) for each combination of simulation factors.

The proportion of significant Q tests across the 2000 replications represented empirical Type I error rates and power values and were used to judge the robustness of the Q test to assumption violations. The resulting 4 (score distribution) x 3 (Sample size) x 3 (Number of studies) x 4 (δ values) design was replicated to permit error variation to be estimated within each cell. Thus, two empirical empirical proportions of rejections per cell were generated.

Results

Adequacy of the Simulation

The adequacy of the simulation was judged by examining the skewness, kurtosis, and d_i values across the conditions studied, and by examining empirical Type I error rates and power values when the scores were normally distributed for large N_i, in which case these values should be close to theoretical values. After examining plots of the simulated scores, skewness and kurtosis indices were computed. The normal approximation was quite good, producing skewness and kurtosis values very close 0. The nonnormal distributions all showed the pattern of producing skewness and kurtosis values equal to or slightly less than the specified γ_1 and γ_2 values. For example, for γ_1 = 1.5, γ_2 = 5, the average skewness and kurtosis values were 1.4 and 4.85, repectively; for γ_1 = 2, γ_2 = 6 the average values values were 1.9 and 5.8, respectively. Thus,

the simulated nonnormal data were slightly less nonnormal than anticipated. The d_i were also quite close to the target values, especially for larger sample sizes. Even for $N_i = 10$ the deviation of the d_i from the specified value tended to be modest. For example, for $\delta = 1$, $N_i = 10$, and a normal distribution, the average d_i was .95. On the whole, the simulated data appeared to possess (approximately) the desired properties.

Type I Error Rates of the Q Test

When δ equaled 0 the proportion of rejections represented empirical error rates. These values are reported in Table 1. Because the empirical error rates for $\alpha = .01$, .05, and .10 produced similar patterns, only the values associated with .05 appear in Table 1. Perhaps the most striking feature of the $\delta = 0$ results is that almost all of the Type I error rates are below .05 and that many are quite conservative, especially for larger k paired with a smaller N_i. The k = 5 results are consistent with those reported by Hedges and Olkin (1985, p. 125) but conflict with those of Chang (1992), who reported inflated Type I error rates as large as .10 for the large k, small N_i pairing. A few additional computer runs were done with $N_i = 120$ (60 per group) to see if empirical error rates converged to .05. For k = 5, 10, and 30, the error rates for $N_i = 60$ were .039, .056, and .053, respectively, the latter two being within an acceptable range if sampling error is taken into account. The .039, on the other hand, was still conservative. Type of distribution appeared to have little effect on error rates. On the other hand, k and N_i appeared to have a direct effect on error rates.

Power of the Q Test

Setting $\delta = .5$, 1, or 1.5 produced estimated power values for the Q test. These values are also reported in Table 1, where the resulting pattern is similar to that observed in the Type I

error case; namely, power was largest for a given δ when larger N_i were paired with smaller k, and decreased as k increased for a fixed N_i, a result which agrees with Chang (1992). In all cases, N_i and k appeared to be the dominant factors, whereas the effect of increasingly nonnormal dstributions appeared to be to slight. Predictions about power appear to depend heavily on the relationship between N_i, k, and δ.

Theoretical power values were computed to assess their agreement with empirical values by assuming a normal distribution for the scores and using the equation developed in Chang (1992) and the noncentral chi-square table in Owen (1962). Theoretical power values for the δ = 1.5 case and α = .05 are illustrative of the general pattern and are reported in Table 1 in parentheses.

The comparison of empirical versus theoretical power values for all values of k and N_i suggest two conclusions. First, empirical power values decreased dramatically as k increased, especially for smaller N_i, so much so that in some cases the power was only slightly larger than the empirical Type I error rate (Overall, 1969 discusses this phenomenon). Second, the empirical and theoretical power values in Table 1 tend to agree with Chang's findings that the magnitude of the misfit depends heavily on how k and N_i are paired and that discrepancies shrink as N_i increases, but the two sets of findings disagree in the direction of the misfitting. The results in Table 1 indicate that empirical power values were typically less than theoretical values for the small N_i, large k cases, whereas Chang reported that empirical power values typically overestimated theoretical values for these conditions. This discrepancy may be attributable to different patterns of Type I error rates for the large k, small N_i pairing. Chang's inflated Type I error rates under these conditions would, other things being equal, be expected to produce higher power values, whereas the conservative Type I error rates reported in Table 1 could

explain the underestimation of power. The results in Table 1 are generally consistent with Chang's findings for larger N_i.

Data Analysis

Examining Table 1 is instructive but leaves open the possibility that important patterns in the empirical proportions of rejections may be missed or that the magnitude of effects may be missestimated. To test for the presence of interactions and to estimate the magnitude of significant effects the empirical proportions were analyzed using weighted least squares multiple regression. The predictors in these models were N_i, k, γ_1, and γ_2, the latter two variables being used to represent type of distribution.[2] The predictors were centered to minimize collinearity problems due to scaling. The proportions of rejections (e.g., $\hat{p}$) served as outcomes, with weights of $(\hat{\sigma}^2_{\hat{p}})^{-1}$. Analyses were conducted separately for the $\delta = 0$ (Type I error) and $\delta \neq 0$ (power) cases. Only the results for the $\alpha = .05$ case are reported in Table 2.

Two regression models were fitted to the empirical error rates: a main effects model and a second model containing both main effects and two-way interaction terms. This allowed the contribution of the interactions to be investigated. An examination of the residuals revealed no unusual patterns in the data.

The results in Table 2 for $\delta = 0$ indicate that the empirical error rates of the Q test were insensitive to the predictors. This supports the notion that the Type I error rate of the Q test is generally robust, although it is worth restating that the error rates were uniformly below .05. The model 2a results for the $\delta \neq 0$ case indicate that the empirical power values proved to be quite sensitive to the predictors in models 2a and 2b. The $R^2_{adj} = .98$ for model 2b indicates that

[2]Chang (1992) used $N^{1/2}$ and $K^{1/2}$ as predictors. Analyses were done using N and K and, separately, $N^{1/2}$ and $K^{1/2}$. These results were similar.

virtually all of the variation in the power values is explained by the regression model. The different R^2_{adj} values in model 1 versus model 2 occur because, while the Type I error rate of a statistical test like the Q test may be insensitive to factors like type of score distribution, its power is directly and highly dependent on noncentrality parameters. Similar results have been reported by Harwell, et al., (1992) and Lix, Keselman, and Keselman (1992). Restricting the predictor values of k to 10 and 30, and those of N_i to 10 and 20, which seemed to have the greatest effect on Type I errors and power, produced regression results very similar to those in Table 2. Thus, the conclusions do not appear to hold only for the large k, small N_i pairing. Interestingly, all of the estimated standardized regression coefficients for model 2 were less than .06 in value and fairly indistinguishable.

Conclusions

It appears that meta-analysts need not be concerned that nonnormaly score distributions will have much effect on Type I or Type II error rates of the Q test. However, the pairing of study sample size and number of studies appear to play a crucial role in the Type I and Type II error behavior of the Q test. Chang (1992) summarized her findings by statuing put it, "...homogeneity tests were more sensitive than indicated by theory for data with small sample sizes..." (p. 59). The findings of the present study and those of Chang (1992) support this statement, but disagree in the direction of the sensitivity. Both sets of findings suggest the Type II error rate of the Q test is affected by particular pairings of study sample size and number of studies, but disagree in whether the probability of a Type II error is higher or lower than indicated by theory. On the other hand, for pairings in which study sample size is noticeably larger than the number of studies in the meta-analysis, both sets of findings agree that the likelihood of committing a Type II error with the Q test is consistent with theory.

Implications for Future Research

Additional empirical studies are needed to resolve current discrepancies in the behavior of the Type I and Type II error rates of the Q test for specific pairing of study sample size and number of studies, and to provide evidence about the magnitude of the discrepancies. Another useful addition to the metaanalytic literature would tables of noncentrality values for Q for combinations of study sample size and number of studies.

References

Alexander, R.A., Scozzaro, M.J., & Borodkin, L.J. (1989). Statistical and empirical examination of the chi-square test of homogeneity of correlations in meta-analysis. **Psychological Bulletin, 106,** 329-331.

Bailey, K.R. (1987). Inter-study differences: How should they influence the interpretation and analysis of results? **Statistical Medicine, 6,** 351-358.

Box, G.E.P. (1953). Non-normality and tests on variances. **Biometrika, 40,** 318-334.

Chang, L. (1992). **A power analysis of the test of homogeneity in effect-size meta-analysis**. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Fleishman, A. (1978). A method for simulating nonnormal distributions. **Psychometrika, 43,** 521-532.

Fleiss, J.L., & Gross, A.J. (1991). Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. **Journal of Clinical Epidemiology, 44,** 127-139.

Gayen, A.K. (1949). The distribution of 'Student' t in the random samples of any size drawn from non-normal universes. **Biometrika, 36,** 353-369.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper and L. V. Hedges (Eds.), **The Handbook of Research Synthesis**. New York: Sage.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. **Journal of Educational Statistics, 17,** 315-339.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. **Journal of Educational Statistics, 6,** 107-128.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. **Journal of Educational Statistics, 7,** 119-137.

Hedges, L. V., & Olkin, I. (1985). **Statistical methods for meta-analysis.** New York: Academic Press.

Hoaglin, D.C., & Andrews, D.F. (1975). The reporting of computation-based results in statistics. **The American Statistician, 29,** 122-126.

Lewis, P.A.W., & Orav, E.J. (1989). **Simulation methodology for statisticians, operations analysts, and engineers**. (Vol. I). Pacifica Grove, CA: Wadsworth and Brooks/Cole.

Naylor, T.H., Balintfy, J.L., Burdick, D.S., & Chu, K. (1968). **Computer simulation techniques**. New York: Wiley.

Overall, J.E. (1969).Classical statistical hypothesis testing within the context of Bayesian theory. **Psychological Bulletin**, **71**, 285-292.

Owen, D.B. (1962). **Handbook of statistical tables**. Reading, MA: Addison-Wesley.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1986). **Numerical recipes**. Boston, MA: Cambridge University Press.

Rosenthal, R., & Rubin, D.B. (1982). Comparing effect sizes of independent studies. **Psychological Bulletin**, **92**, 500-504.

Sawilosky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality, **Psychological Bulletin, 3**, 352-360.

Shadish, W.R., & Haddock, C.K. (1994). Combining estimates of effect size. In H. Cooper & L.V. Hedges (Eds.), **The Handbook of Research Synthesis**. New York: Sage.

Srivastava, A.B.L. (1959). Effects of non-normality on the power of the analysis of variance test. **Biometrics**, **46**, 112-122.

Wolf, F. M. (1990). Methodological observations on bias. In K. W. Wachter and M. L. Straf (Eds.), **The future of meta-analysis**. New York: Sage.

Table 2+
Empirical Results for the Q Test

		N_i		10			20			40	
Distribution	δ	K	5	10	30	5	10	30	5	10	30
	0		042	029	022	041	049	031	046	050	040
$\gamma_1 = \gamma_2 = 0$	.5		061	038	026	105	065	047	153	120	074
Normal	1		111	087	043	265	185	103	551	441	230
	1.5		218(.29)	140(.24)	060(.15)	535(.58)	400(.45)	195(.28)	902(.88)	827(.82)	574(.62)
	0		032	031	016	035	035	027	041	037	039
$\gamma_1 = 1, \gamma_2 = 3$	.5		055	049	028	092	061	046	160	134	078
	1		130	083	036	303	190	089	580	474	261
	1.5		239	154	057	590	455	207	915	865	555
	0		029	026	019	041	032	032	049	041	028
$\gamma_1 = 1.5, \gamma_2 = 5$	.5		080	038	016	099	068	038	157	126	073
	1		137	082	036	291	206	099	580	474	274
	1.5		272	170	056	588	463	219	900	826	586
	0		035	024	015	034	032	019	039	037	031
$\gamma_1 = 2, \gamma_2 = 6$	.5		057	041	018	094	070	036	170	138	080
	1		146	091	038	311	213	107	579	468	240
	1.5		297	182	065	608	477	213	893	834	608

+Note: Each of the tabled values is the average of two empirical proportions of rejections, K = number of studies, N_i = study sample size, γ_1 = skewness, γ_2 = kurtosis. Values in parentheses are theoretical power values.

Table 2*
Analysis of Empirical Error Rates

$\delta = 0$

Model 1	$df_{Regression}$	$df_{Residual}$	R^2_{adj}
1a	4	67	not sig.
1b	10	61	not sig.

$\delta \neq 0$

Model 2	$df_{Regression}$	$df_{Residual}$	R^2_{adj}
2a	4	211	.42
2b	14	201	.98

*Note. Model 1a and 1b were main effects models which used the skewness, kurtosis, study sample size, and the number of studies as predictors; models 1b and 2b used both main effects and two-way interactions as predictors.